IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re the Application of:

Philip Victor HARMAN et al

Serial No.: 09/925,932

Filed: August 9, 2001

For: IMAGE CONVERSION AND ENCODING TECHNIQUES

Atty. Docket No.: 006020.00012

CLAIM FOR PRIORITY UNDER 35 U.S.C. §119

Commissioner for Patents
Washington, D. C. 20231

Sir:

The benefit of the filing date of the following prior foreign application is hereby requested for the above-identified application and the priority provided under 35 U.S.C. §119 is hereby claimed: (a certified copy of each foreign application is enclosed herewith)

Country	Application Number	Date of Filing (mm-dd-yyyy)
Australia	PQ9292	9 August 2000
Australia	PR0455	29 September 2000

It is requested that the file of this application be marked to indicate that the requirements of 35 U.S.C. §119 have been fulfilled and that the Patent and Trademark Office kindly acknowledge receipt of these documents.

Respectfully submitted,

BANNER & WITCOFF, LTD.

Dated: December 6, 2001

By: Gary D. Fedorochko

Gary D. Fedorochko
Registration No. 35,509

1001 G Street, N.W.
Washington, D.C. 20001-4597
(202) 508-9100

GDF:lab

AUSTRALIA

Patent Office
Canberra

CERTIFIED COPY OF PRIORITY DOCUMENT

I, GAYE TURNER, TEAM LEADER EXAMINATION SUPPORT AND SALES hereby certify that annexed is a true copy of the Provisional specification in connection with Application No. PQ 9292 for a patent by DYNAMIC DIGITAL DEPTH RESEARCH PTY LTD filed on 09 August 2000.

COMMONWEALTH OF AUSTRALIA
AUSTRALIA
PATENT OFFICE

WITNESS my hand this
Fourteenth day of August 2001

GAYE TURNER
TEAM LEADER EXAMINATION
SUPPORT AND SALES

P/00/009 28/5/91
Regulation 3.2

AUSTRALIA

Patents Act 1990

ORIGINAL

PROVISIONAL SPECIFICATION

IMAGE CONVERSION AND ENCODING TECHNIQUES

The invention is described in the following statement:

IMAGE CONVERSION AND ENCODING TECHNIQUES

FIELD OF INVENTION

The present invention is generally directed towards stereoscopic image synthesis and more particularly toward an improved method of converting two dimensional (2D) images for further encoding, transmission and decoding for the purpose of stereoscopic display.

BACKGROUND

The applicants have previously described in PCT/AU96/00820, a method of producing left and right eye images for a stereoscopic display from an original 2D image including the steps of:

(a) identifying at least one object within the original image;

(b) outlining each object;

(c) defining a depth characteristic for each object; and

(d) respectively displacing selected areas of each object by a determined amount in a lateral direction as a function of the depth characteristic of each object, to form two stretched images for viewing by the left and right eyes of the viewer.

These steps can be individually and collectively referred to as Dynamic Depth Cueing or DDC.

Additionally, the Applicants have previously described in PCT/AU98/0155 in one aspect a method of encoding a depth map including the steps of:

(a) allocating an object number to an object;

(b) allocating the object with a depth; and

(c) defining the object outline.

The object outline may be defined by a series of co-ordinates, and/or curves.

In another aspect the previous invention disclosed the use of curves to generate an outline of an object in a 2D to 3D conversion process.

In a further aspect the previous invention disclosed the use of curves to define an object in a 2D to 3D conversion process.

In a further aspect the previous invention disclosed a method of transmission of depth map information wherein the information is included in the Vertical Blanking Interval or MPEG data stream.

In a further aspect the previous invention disclosed the use of generic libraries to assist in the 2D to 3D conversion process.

The present invention further improves on the operation of the Applicant's earlier systems.

Additionally, the Applicants have previously described in (****DDC 3) in one aspect a method of encoding a depth map including:

allocating an object identification symbol to an object;

using the allocated object symbol to represent the shape of the object;

allocating the object with a depth;

compressing the information representing the object and its depth;

transmitting and/ or storing this information ; and

decompressing the information.

OBJECT OF THE INVENTION

It is the object of this invention to describe a technique to improve the previously disclosed processes for converting 2D images into stereoscopic 3D.

SUMMARY OF THE INVENTION

The Applicant's previous disclosures have related to techniques for converting 2D images into stereoscopic 3D images. The conversion processes disclosed incorporated the generation of a depth map that was associated with a 2D image. In one embodiment the depth maps were created manually on a frame by frame basis. The improvement described in this application enables a fewer number of key-frames to have depth maps created and the intermediate depth maps calculated. Since the key-frames represent a small fraction of the total number of frames, this new technique represents a substantial improvement in conversion efficiency both in terms of time and cost.

BRIEF DESCRIPTION OF THE INVENTION

The invention is intended to improve the process of producing depth maps for associated 2D images. The invention requires 2D images and associated depth maps to be provided at selected key-frames. The depth maps at these key-frames may be manually generated as previously disclosed by the applicants

or produced automatically using depth capture techniques including, although not limited to, laser range finders i.e. LIDAR (LIght Direction And Range) devices and depth-from-focus techniques.

The 2D image and associated depth map, for each key-frame, is presented to an algorithm that is capable of learning the relationship between the depth z assigned to each pixel in the image, its x and y location and image characteristics. The image characteristics include, although not limited to, the RGB value of each pixel. In general the algorithm solves the equation

$$z = f(x,y,R,G,B)$$

for each pixel in the key-frames.

The algorithm is then presented with each subsequent frame between the adjacent key-frames and for each pixel uses the algorithm to calculate the value of z.

IN THE DRAWINGS

Figure 1 shows a flow chart representing the training process.

Figure 2 shows a flow chart representing the conversion process.

DETAILED DESCRIPTION OF THE INVENTION

The invention provides a technique for constructing the 3D stereo information from a sequence of 2D images. From illustration purposes only it is assumed that the 3D stereo information is a depth map. However, it is not intended to limit the claims of this disclosure to depth maps since the approach can also be applied to other stereoscopic information, for example disparity information obtained from a key- frame comprising a stereo pair.

PROCESS

SELECT KEY-FRAMES

The key-frames are chosen manually. For these key-frames, the depth maps are already available. This availability may be due to any process, such as, but not limited to, human specification or some stereo information capture process for example a LIDAR device. For all other frames in the 2D image sequence, the invention provides specification of the depth maps, based on the key-frame information initially available. It is expected that the number of key-frames will be a small fraction of the total number of frames. Hence the invention

provides a way of vastly reducing the amount of depth maps required to be initially generated.

CREATE MAPPING FUNCTION

The invention observes the key-frames and the corresponding depth map initially available, in order to create a mapping function. The mapping function may be a process which takes as input any given measurement of a 2D image, and provides as output a depth map for that image. This mapping is learnt by capturing the relationship between the key-frame image data and depth map available for those images.

The mapping function may take the form of any generic processing unit, where input data is received, processed, and an output given. Preferably, this processing unit is amenable to a learning process, where its nature is determined by examination of the key-frame data, and its corresponding depth map. In the field of machine learning, such mapping functions are well known and include, although not limited to, neural networks, decision trees, decision graphs, model trees and nearest-neighbour classifiers.

LEARN RELATIONSHIPS BETWEEN INPUT DATA AND DESIRED OUTPUT DATA

In a learning process, information from the 2D key-frame image is presented to the mapping function. This information may be presented on a pixel by pixel basis, where pixel measurements are provided, such as red, green and blue values, or other measurements such as luminance, chrominance, contrast and spatial measurements such as horizontal and vertical positioning in the image. Alternatively, the information may be presented in the form of higher level image features, such as larger sets of pixels and measurements on a set of pixels such as mean and variance or edges, corners etc (i.e. the response of a feature detector). Larger sets of pixels may for example represent segments in the image, being sets of connected pixels forming a homogenous region.

For illustrative purposes only, the 2D image may be represented in the form

x,y,R,G,B

where x and y represent the x and y coordinates of each pixel and R,G,B represent the red, green and blue value of that pixel.

Next, the corresponding depth map is presented to the mapping function, so that it may learn its required mapping. Normally individual pixels are presented to the mapping function, however, if higher level image features are being used, such as larger sets of pixels, or segments, the depth map may be a measurement of the depth for that set of pixels, such as mean and variance.

For illustrative purposes only, the depth map may be represented in the form

z,x,y

where x and y represent the x and y coordinates of each pixel and z represents the depth value assigned to that corresponding pixel.

The process of learning this relationship between input data, and desired output data is well understood by those skilled in the art of artificial intelligence, and may take on many forms. Preferred embodiments of a learning algorithm, are those that seek to design a mapping function which minimises some measurement of mapping error.

The learning algorithm attempts to generalise the relationships between the 2D image information and the depth map present in the key-frame examples. This generalisation will then be applied to complete the depth maps for the entire sequence. Examples of successful learning algorithms known in the art are the back-propagation algorithm for learning neural networks, the C4.5 algorithm for learning decision trees, and the K-Means algorithm for learning cluster-type classifiers.

For illustrative purposes only, and to aid understanding, the learning algorithm may be considered to compute the following relationship for each pixel in the 2D image

$$z_n = k_a.x_n + k_b.y_n + k_c.R_n + k_d.G_n + k_e.B_n$$

where

n is the nth pixel in the key-frame image

z_n is the value of the depth assigned to the pixel at x_n,y_n

k_a to k_e are constants and are determined by the algorithm

R_n is the value of the Red component of the pixel at x_n,y_n

G_n is the value of the Green component of the pixel at x_n,y_n

B_n is the value of the Blue component of the pixel at x_n,y_n

It will be appreciated by those skilled in the art that the above equation is a simplification for purposes of explanation only and would not work in practice. In a practical implementation, using for example a neural network and given the large number of pixels in an image, the network would learn one large equation containing may k values, multiplications and additions.

This process is illustrated in Figure 1.

APPLY MAPPING FUNCTION TO 2D IMAGE

The invention next takes this mapping function and applies it across a set of 2D images that do not yet have depth maps available. For a given 2D image in that set, the inputs to the mapping function are determined in a similar manner as that presented to the mapping function during the learning process. For example, if the mapping function was learnt by presenting the measurements of a single pixel as input, the mapping function will now require these same measurements for the pixels in the new image. With these inputs, the mapping function performs its learnt task and outputs a depth measurement. Again, in the example for a single pixel, this depth measurement may be a simple depth value. In this example, the mapping function is applied across the entire image, to complete a full set of depth data for the image. Alternatively, if the mapping function was trained using larger sets of pixels, it is now required to generate such larger sets of pixels for the new image. The higher-level measurements on these sets of pixels are made, such as mean and variance, in the same manner as that during the learning process. With these inputs now established, the mapping function produces the required depth measurement, for that set of pixels.

For illustrative purposes only, and to aid understanding, the algorithm determines the depth, z_n, at each pixel in the 2D image by applying the following relationship

$$z_n = k_a.x_n + k_b.y_n + k_c.R_n + k_d.G_n + k_e.B_n$$

where

n is the nth pixel in the image

z_n is the value of the depth assigned to the pixel at x_n,y_n

k_a to k_e are constants previously determined by the algorithm

R_n is the value of the Red component of the pixel at x_n,y_n

G_n is the value of the Green component of the pixel at x_n,y_n

B_n is the value of the Blue component of the pixel at x_n,y_n

For a sequence of 2D images, key-frames with depth maps may be spaced throughout the sequence, in any arbitrary way. In the preferred embodiment, the mapping function will be presented with a set of key-frames, and their corresponding depth maps, which span a set of 2D images that have some commonality. In the simplest case, two key-frames are used to train the mapping function, and the mapping function then used to determine the depth maps for the 2D images between the two said key-frames. However, there is no restriction to the number of key-frames which may be used to train a mapping function. Further, there is no restriction to the number of mapping functions that are used to complete a full set of 2D images. For each pair of adjacent key-frames, a new mapping function may be used.

This process is illustrated in Figure 2.

ALTERNATIVE EMBODIMENTS

In an alternative embodiment, the mapping function will be presented with a larger number of key-frames for training, with an added training variable representing the passage of time through the image sequence. Referring to the previous example, the depth value, z, for a given pixel is calculated as a function of the form:

$$z_n = k_a.x_n + k_b.y_n + k_c.R_n + k_d.G_n + k_e.B_n$$

The time variable is now introduced such that this function is extend to read:

$$z_n = k_a.x_n + k_b.y_n + k_c.R_n + k_d.G_n + k_e.B_n + k_f.T$$

where:

n is the nth pixel in the image

z_n is the value of the depth assigned to the pixel at x_n,y_n

k_a to k_f are constants previously determined by the algorithm

R_n is the value of the Red component of the pixel at x_n,y_n

G_n is the value of the Green component of the pixel at x_n,y_n

B_n is the value of the Blue component of the pixel at x_n,y_n

T is a measurement of time, for this particular frame in the sequence

This value of time may be set to zero for the starting frame in the sequence, and set to 1.0 for the final frame in the sequence. All frames in between naturally have a time value representing their relative progress towards the final frame. When training the mapping function, the time value for the key-frames is calculated and presented to the mapping function so that the learning algorithm may incorporate this information. When the mapping function is used to convert the remaining non key-frames in the sequence, for each frame its time value is calculated, and used as input in the mapping function. This mapping function is used, as described previously, to calculate the depth information for that image.

The addition of this time variable assists the training function in generalising the information available in the key-frames. In the absence of a time variable, it is possible that the depth information in two key-frames may contradict each other. This might occur when pixels of a similar colour occur in the same spatial region in both key-frames, but belong to different objects. For example, in the first key-frame, a green car may be observed in the centre of the image, with a depth characteristic bringing it to the foreground. In the next key-frame, the car may have moved, revealing behind it a green paddock, whose depth characteristic specifies a middle ground region. The training algorithm is presented with two key-frames, that both have green pixels in the centre of the image, but have different depth characteristics. It will not be possible to resolve this conflict, and the mapping function is not expected to perform well in such a region. With the introduction of a time variable, the training algorithm will be able to resolve the conflict by recognising that the green pixels in the centre of the image, are foreground pixels at a time near the first key-frame in the image sequence. As time progresses towards the second key-frame, the training algorithm will become more inclined to recognise green pixels in the centre of the image as the middle-ground depth of the green paddock.

The addition of the time variable also enables the algorithm to be trained with multiple key-frames. Without the time variable the key-frames may contain contradictions, due for example to new objects appearing, over a long sequence. The addition of a time variable enables the learning algorithm to perform correctly

over multiple key-frames since it enables the algorithm to account for the changes in the key-frames.

For example, assuming the process is being used to calculate depth maps from a video sequence at 50 frames per second. From this sequence six key-frames, each comprising a 2D image and associated depth map, spaced 10 frames apart have been selected. The algorithm will be presented with these six key-frames, the time variable being assigned as follows:

Key-frame number	0	10	20	30	40	50
Time variable	0	0.2	0.4	0.6	0.8	1.0

These values of time would be presented to the learning algorithm for the each pixel in the corresponding key-frame.

When the algorithm is interpolating the z data for frame 14 we present to the learning algorithm, for each pixel in frame 14, a time value of

$(14/50) = 0.28$

from this value the algorithm can determine that the frame to be calculated is in the region of key frames 10 and 20 and calculate accordingly.

In a further alternative embodiment, the training algorithm may attempt to introduce a random component into the key-frame information. With any learning algorithm this overcomes the difficulty of over-training. Over-training refers to the situation where the learning algorithm simply remembers the key-frame training information. This is analogous to a child wrote-learning multiplication tables, without gaining any understanding of the concept of multiplication itself. This problem is well-known in the field of machine learning, and a common approach to relieving the problem is to introduce random noise into the training data. A good learning algorithm will be forced to distinguish between the noise in the training data, and the quality information. In doing this, it will be encouraged to learn the nature of the data rather than simply remember it. An example embodiment of this approach refers to the previous example, where the training algorithm learns the function:

$$z_n = k_a.x_n + k_b.y_n + k_c.R_n + k_d.G_n + k_e.B_n$$

When presenting the inputs to the training algorithm, being x,y,R,G and B, a small noise component is added to these values. The noise component may be a small positive or negative random number.

An alternative embodiment may exploit the fact that the mapping functions give a full representation of the depth information for all non key-frame images in the sequence. The mapping function could be viewed as an encoding of this depth information. It is expected that the mapping function may be transmitted with a relatively small amount of data, and hence represents a significant compression of the depth information.

Consider the case where there are two key-frames, 20 frames apart in the sequence. A mapping function has been learnt for these two key-frames, and this mapping function now provides all depth information for the intermediate frames. The mapping function itself represents a compression of all this depth information across the twenty frames. If, for example purposes only, the mapping function can be written to a file using 6000 bytes, then for this price we gain 20 frames worth of depth information. Effectively, this represents a file size of 6000 / 20 = 300 bytes per frame. In a practical implementation the effective compression will be substantial.

In a further embodiment, this above compression may allow for efficient transmission of 3D information, embedded in a 2D image source i.e. a 2D compatible 3D image. Since the mapping functions require a file length that is typically a tiny fraction of the 2D image data that it provides 3D information for, the addition of 3D information to the 2D image sequence is achieved with a very small overhead.

In this case, the 3D information is generated prior to viewing, or in real-time, at the viewing end, by simply applying the mapping function over each 2D image in the sequence as it is viewed. This is made possible by the fact that the types of mapping functions found in machine learning are very efficient in providing calculations *after* they have been trained. Typically the training process is slow and resource intensive, and is usually performed offline during the process of building the 3D image content. Once trained, the mapping function may be transmitted to the viewer end and will perform with a very high throughput suitable for realtime conversion of the 2D image to 3D.

OTHER APPLICATIONS

It is a specific intent of this disclosure that the invention should be applied to the creation of depth maps for other than the production of stereoscopic images.

It will be known to those skilled in the art that depth maps are used extensively within the special effects industry. In order to compose live action, or computer generated images, within a 2D image it is frequently necessary to manually produce a depth map for each frame of 2D image. These depth maps enable the additional images to be composed so as to appear to move with the appropriate geometry within the original 2D image.

It is also know that cameras are being developed that enable a depth map to be obtained from a live image. Typically these use laser range finding techniques and are generically know as LIDAR devices. In order to capture depth maps at television frame rates an expensive and complex system is required. The application of this invention would enable simpler and less complex LIDAR devices to be constructed that need only capture depth maps at a fraction of the video field rate, or other infrequent periods, and the missing depth maps produced by interpolation using the techniques described in this invention.

DATED this 9th day of August 2000

DYNAMIC DIGITAL DEPTH RESEARCH PTY LTD

WATERMARK PATENT & TRADEMARK ATTORNEYS
4TH FLOOR DURACK CENTRE
263 ADELAIDE TERRACE
PERTH WA 6000

Figure 1 - Training Process

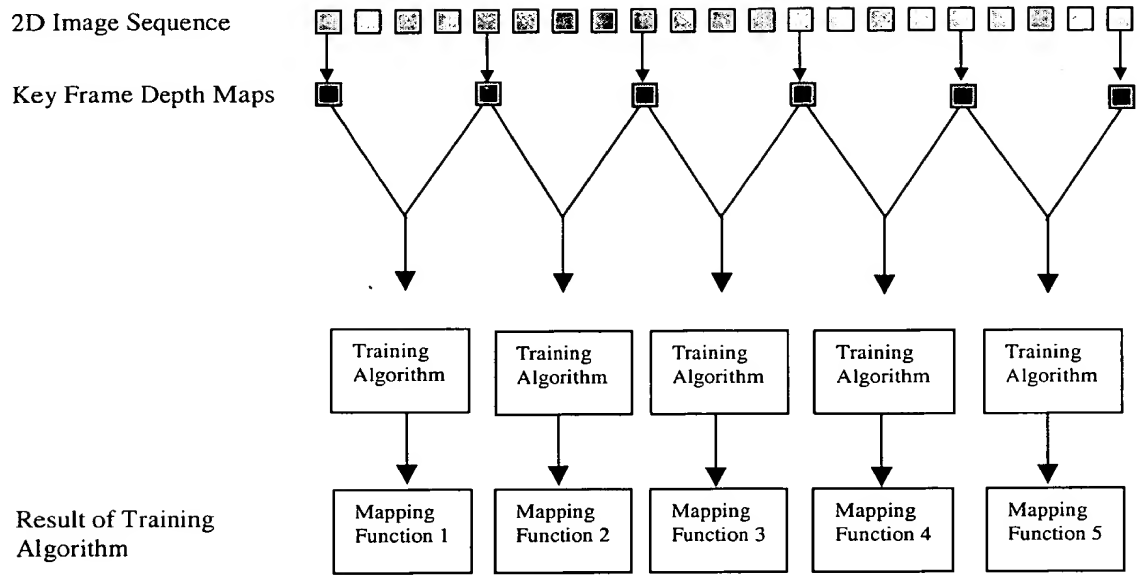

Figure 2 - Conversion Process